Assembling a physical map of the genome by marker sequences

ZHANG Pei-yu (张培玉)*, ZHANG Hong-hai (张洪海)
(*Department of Biology, Qufu Normal University, Qufu City 273165, Shandong Province, P.R. China*)

HUA Yu-ping (华育平)
(*College of Wildlife Resource, Northeast Forestry University, Harbin 150040, P.R. China*)

XU Lai-xiang (徐来祥)
(*Department of Biology, Qufu Normal University, Qufu City 273165, Shandong Province, P.R. China*)

Abstract: Molecular genetic maps were commonly constructed by analyzing the segregation of restriction fragment length polymorphisms (RFLPs). Here we described methodology-marker sequences in a new mapping based on recent documents. With the methods they were unique sequences detected by the polymerase chain reaction (PCR). Each of the methods had its limitations and the current trend was to integrate the maps produced by the different methods. Marker sequences contained mainly expressed sequence tags (ESTs), polymorphic sequence-tagged sites (STSs), randomly amplified polymorphic DNA (RAPDs), cleaved amplified polymorphic sequences (CAPS), amplified fragment length polymorphism (AFLPs), genome sequence sampling (GSS) and sequence-tagged connectors (STCs) in this paper.

Key words: Marker sequences; Sequence-tagged sites; Expressed sequence tags; Randomly amplified polymorphic DNA; Cleaved amplified polymorphic sequences; Amplified fragment length polymorphism; Genome sequence sampling; Sequence-tagged connectors

CLC number: Q75 **Document code:** A **Article ID**: 1007-662X(2000)02-0127-05

Introduction

After a genome has been fragmented, the fragments cloned to generate a genomic library. It is necessary to assemble the cloned fragments in the same linear order as found in the chromosomes from which they are derived. One method of linking cloned fragments was chromosome walking (Bender *et al.*1983), which was originally developed for the isolation of gene sequences whose function was unknown but whose genetic location was known. For the purposes of mapping, a single cloned fragment was selected and used as a probe to detect other clones in the library, with which the probe can hybridize. The negative hybridization signals represent clones overlapping with the same sequence as the probe. The overlap can be to the right or the left. Thus the single walking step can occur in both directions along the chromosome if it is repeated many times.

Although inherently attractive, chromosome walking is too laborious and time consuming to be of value for mapping most genomes. In practice, many more 'walks' would be necessary since a representative library would contain a much larger number of clones. Three general methodologies for mapping have been developed as alternatives to chromosome walking: restriction enzyme fingerprinting, marker sequences and hybridization assays. In each case the objective is to create a landmark map of the genome under study with marks dispersed at regular intervals throughout the map. In restriction enzyme fingerprints the markers are restriction sites; in the other two methods the markers are unique sequences detected by the polymerase chain reaction or by hybridization. Each of the methods has its limitations and the current trend is to integrate the maps produced by the different methods. In some cases the generation of a map requires the use of all three methods, as was found by Chumakov *et al.* (1995) in the construction of a human YAC contig map. Here we described only the new mapping methodology using marker sequences.

Types and characteristics of marker sequences

Sequence-tagged sites (STSs).

The concept of sequence-tagged sites (STSs) was developed by Olson *et al.* (1989) in an attempt to systematize landmarking of human genome. Ba-

Foundation item: The project was supported by National Science Foundation of China (No.39970103).

Biography: *ZHANG Pei-yu (1964-), male, lecture of Department of Biology, Qufu Normal University, Qufu City 273165, Shandong Province, P.R. China.

Received date: 2000-01-18

Responsible editor: Zhu Hong

sically, the STS is a short region of DNA about 200-300 bases long whose exact sequence is found nowhere else in the genome. Two or more clones containing the same STS must overlap and the overlap must include the STS.

Any clone that can be sequenced may be used as a STS provided it contains a unique sequence. A better method to develop STS markers is to create a chromosome-specific library in phage M_{13}. Random M_{13} clones are selected and 200-400 bases sequenced. The sequence data generated are compared with all known repeated sequences to help identify regions likely to be unique. Two PCR primer sequences are selected from the unique regions which are separated by 100-300 bp and whose melting temperatures are similar. Once identified, the primers are synthesized and used to PCR amplify genomic DNA from the target organism and the amplification products analyzed by agarose gel electrophoresis. A functional STS marker will amplify a single target region of the genome and produce a single band on the electrophoretic gel at a position corresponding to the size of the target region. Alternatively, a STS marker can be used as a hybridization probe.

Operationally, a STS is specified by the sequence of the two primers that make its production possible. Thus once it is defined, it could be applied to all despite, whether the region in the study is cloned in a phage, a cosmid, a BAC, PAC or YAC. Moreover, the STS will remain validity if the corresponding area of the genome is re-cloned sometime in the future. The STS is fully portable once the sequences of two primers are known and available from databanks.

The principle of the use of STSs to generate physical maps has been confirmed by a number of works. Thirty YAC clones from the cystic fibrosis region of human chromosome 7 were assembled into a single contig (Green & Olsen 1990) that spans more than 1.5 Mb. At the same time, individual YACs as large as 790 kb and containing the entire cystic fibrosis gene were constructed *in vivo* by meiotic recombination in yeast of overlapping YACs. Foote *et al.* (1992) were able to assemble 196 recombinant clones into a single overlapping array, which included over 98% of the enchromatic portion of the human Y chromosome. Similarly, Chumakov *et al.* (1992) were able to assemble to generate a STS map for human chromosome 21q, which was consistent with physical and genetic mapping data.

Expressed sequence tags (ESTs)

In organisms with large amounts of repetitive DNA, the generation of an appropriate sequence, and confirmation that it is a STS, can be time consuming. Adams *et al.* (1991) had suggested an alternative approach. The principle of the method is based on the observation that spliced mRNA contains sequences that are largely free of repetitive DNA. Thus partial cDNA sequences, termed ESTs (expressed sequence tags), can serve the same purpose as the random genomic STSs but have the added advantage of pointing directly to an expressed gene. In a test of this concept, partial DNA sequencing was conducted on 600 randomly selected human cDNA clones to generate ESTs. Of the sequences generated, 337 represented new genes, including 48 with similarity to genes from other organisms, and 36 matched previously sequenced human nuclear genes. Forty-six ESTs were mapped to chromosomes.

In practice, there are a number of operational considerations associated with the use of ESTs. First, they need to be very short to ensure that the two ends of the sequence are contiguous in the genome, i.e. not separated by an intron. Second, large genes may be represented by multiple ESTs, which may correspond to different portions of a transcript or various alternatively spliced transcripts. For example, one of the major databases holds over 1 300 different EST sequences for a single gene product and serum albumin. While this may or may not be a problem in constructing a physical map, it is problematical in the construction of a genetic map.

If it is desirable to select a single representative sequence from each unique gene, then this is accomplished by focusing on 3' untranslated regions (3' UTRs) of mRNAs. This can be achieved using oligo (dT) primers if the mRNA has a poly (A) tail. Two advantages of using the 3' UTRs are that they rarely contain introns and usually display less sequence conservation than coding regions (Makalowski *et al* .1996). The former feature leads to PCR product small enough to amplify and the latter makes it easy to discriminate among gene family members that are very similar in their coding regions.

Polymorphic sequence-tagged sites

So far it has been suggested that a STS yields products with the same size from any DNA samples. However, STSs can also be developed for unique regions along the genome that vary in length from one individual to another. This variation in length most often occurs because of the presence of microsatellites. In man these usually take the form of CA (or GT) repeats with the dinucleotide being repeated 5-50 times. These sequences are very attractive because they are highly polymorphic, i.e.

they will occur as $(CA)_{17}$ in one person, $(CA)_{15}$ in another, and so on. These repeat units are flanked by unique sequences which can act as primers for the generation of the STS. By definition, such STSs are polymorphic and can be traced through families along with other DNA markers.

Polymorphic STSs are particularly useful because, firstly, they occur on average every 10 kb serving as landmarks and secondly, they act as landmarkers on both the physical linkage map and the genetic linkage map for each chromosome and provide points of alignment between the different distance scales on these two types of maps. Weissenbach *et al.* (1992) used a total of 814 of such polymorphic STSs to produce a physical map of the human genome. In 1996 a comprehensive genetic map of the human genome was completed with location of 5 264 polymorphic STSs at 2 335 positions (Dib *et al.* 1996).

Microsatellite markers also have been used to construct a genetic map of the mouse (Dietrich *et al.* 1994, 1996). The map contains 7 377 genetic markers consisting of 6 580 highly informative polymorphic STSs integrated with 797 RFLPs, with an average spacing of 0.2 cM (400 kb). In a similar fashion, polymorphic STS maps have been constructed for pig (Archibald 1994; Archibald *et al.* 1995), cow (Barendse *et al.*1994; Eggen & Fries 1995), sheep (Crawford *et al.*1995) and rat (Serikawa *et al.* 1992; Jacob *et al.*1995). Polymorphic STSs also have been used extensively in plants (Mazur & Tingey 1995) but are being replaced by amplified fragment length polymorphisms (AFLPs).

Randomly amplified polymorphic DNA (RAPDs) and cleaved amplified polymorphic sequences (CAPS)

Polymorphic DNA can be detected by amplification with the absence of the target DNA sequence information used to generate STSs. Williams *et al.* (1990) had described a simple process, distinct from the PCR process, which is based on the amplification of genomic DNA with single primers of arbitrary nucleotide sequence. The nucleotide sequence of each primer was chosen within the constraints that the primer was nine or ten nucleotides in length, between 50 and 80% G+C in composition and contained no palindromes. Not all the sequences amplified in this way are polymorphic but the polymorphic ones can be easily identified. Because of detecting polymorphism of genomic DNA with randomly designed primers, this method is therefore named randomly amplified polymorphic DNA, RAPDs. RAPDs were widely used by plant molecular biologists (Reiter *et al* .1992; Tingey & Del Tufo 1993) to construct maps because they provide very large numbers of markers and are very easy to detect by agrose gel electrophoresis. However, they have two disadvantages. The amplification of a specific sequence is sensitive to PCR conditions, including template concentration, and hence it can be difficult to correlate results obtained by different research groups. For this reason, RAPDs may be converted to STSs after isolation (Kurata *et al.* 1994). A second limitation of the RAPD method is that usually it cannot distinguish heterozygotes from one of the two homozygous genotypes. Nevertheless, Postlethwait *et al.* (1994) had used RAPDs to develop a genetic linkage map of the zebra fish (*Danio rerio*).

A different method for detecting polymorphisms, which is not subject to the problems exhibited by RAPDs, was described by Konieczny and Ausubel (1993). In this method, STSs are derived from genes that have already been mapped and sequenced. Where possible the primers used are chosen such that the PCR products include introns to maximize the possibility of finding polymorphisms. The primary PCR products are subjected to digestion with a panel of restriction endonucleases until a polymorphism is detected. Such markers are called CAPS (cleaved amplified polymorphic sequences). Whereas RFLPs are well suited to mapping newly cloned DNA sequences, they are not convenient to use for mapping genes, such as plant genes, which are first identified by mutation. CAPS are much more useful in this respect.

Amplified fragment length polymorphisms (AFLPs)

AFLP is a diagnostic fingerprinting technique that detects genomic restriction fragments (Vos *et al.* 1995). The major difference is that PCR amplification rather than Southern blotting is used for detection of restriction fragments. The resemblance to the RFLP technique was the basis for choosing the name AFLP. However, the name AFLP should not be used as an acronym because the technique detects presence or absence of restriction fragments but not length differences. The AFLP approach is particularly powerful because it requires no previous sequence characterization of the target genome. For this reason it has been widely used with plant, bacterial and viral genomes (Vos *et al* .1995). It has not proved useful in mapping animal genomes because it is dependent on the presence of high rates of substitutional variation in the DNA; RFLPs are much common in plant genomes compared to animal genomes.

The AFLP technique is based on the amplification of subsets of genomic restriction fragments using PCR. To prepare an AFLP template, genomic

DNA is isolated and digested simultaneously with two restriction endonucleases, *EcoR*I and *MSe*I. The former has a 6 bp recognition site and the latter for a 4 bp recognition site. When used together these enzymes generate small DNA fragments that will amplify well and are in the optimal size range (<1 kb) for separation on denaturing polyacrylamide gels. Following heat inactivation of the restriction enzymes the genomic DNA fragments are ligated to *EcoR*I and *Mse*I adapters to generate template DNA for amplification. These common adapter sequences flanking variable genomic DNA sequences serve as primer binding sites on the restriction fragments. Using this strategy it is possible to amplify many DNA fragments without having prior sequence knowledge.

The PCR is performed in two consecutive reactions. In the first pre-amplification reaction, genomic fragments are amplified with AFLP primers for each having one selective nucleotide. The PCR products of the pre-amplification reaction are diluted and used as a template for the selective amplification using two new AFLP primers that have two or three selective nucleotides. In addition, the *EcoR*I selective primer is radiolabelled. After the selective amplification the PCR products are separated on a gel and the resulting DNA fingerprint detected by autoradiography.

The AFLP technique will generate fingerprints of any DNA regardless of the origin or complexity. The number of amplified fragments is controlled by the cleavage frequency of the rare cutter enzyme and the number of selective bases. In addition, the number of amplified bands may be controlled by the nature of the selective bases. Selective extension with rare di- or trinucleotides will result in a reduction of the number of amplified fragments.

The AFLP technique is not simply a fingerprinting technique. Rather, it is an enabling technology that can bridge the gap between genetic and physical maps. Most AFLP fragments correspond to unique position on the genome and hence can be exploited as landmarks. In higher plants AFLPs may be the most selective way to generate high-density maps. The AFLP markers also can be used to detect corresponding genomic clones. Finally, the technique can be used for fingerprinting of cloned DNA fragments. By using no or few selective nucleotides, restriction fragment fingerprints will be produced which subsequently can be used to line up individual clones and make contigs .

Genome sequence sampling (GSS)

GSS is a recently described technique (Smith *et al.* 1994), which combines elements of restriction fragment mapping and STS mapping and generates maps with a resolution of 1-5 kb. As described above, STSs are used to prepare a physical map of YACs, which ideally have been prepared from isolated chromosomes. To produce a high-resolution map, a chromosome-specific cosmid library is prepared that represents the genome at 20 to 30-fold redundancy and contains a reasonably random distribution of clone ends. This is produced by cloning using a variety of restriction enzymes and cloned sites. Hybridization of YAC probes to the cosmid clones allows selection of large numbers of those included in the YAC clone. The cosmids are restriction mapped and arranged into contigs as described earlier. At the same time the restriction fragments covering the ends of each cloned piece of DNA are identified by hybridization with purified cosmid vector DNA. Next, automated DNA sequence analysis is carried out using cosmid DNA directly as template and primers recognizing each flanking region of the cosmid vector sequence contiguous to the insert. Thus the sequence of 300-500 bp of each end fragment can be determined with limited accuracy and aligned on the map. If a very high density cosmid map has been prepared, it enables a high density sequence map to be prepared. For example, Smith *et al.* (1994) studied the protozoan parasite *Giardia lamblia*, which has a genome size of 10.5 Mb. A cosmid library of 20-fold redundancy would consist of 5 000 unique cosmids and this would generate 10 000 end sequences. Assuming the restriction sites used for cloning were evenly spaced, the DNA sequences determined would be spaced every kilobase, on average.

Sequence-tagged connectors (STCs)

. As noted earlier, low resolution physical maps can be constructed by using STSs or similar means to order YACs. High resolution maps then can be prepared by randomly cutting and sub-cloning YAC inserts into cosmids which then have to be ordered. Venter *et al.* (1996) have proposed an alternative method involving BACs. A BAC library is constructed with an average insert size of 150 kb and a 15-fold coverage of the genome. In the case of the human genome this would require 300 000 clones which would be arranged in microtitre wells. Both ends of each BAC insert are then sequenced for 500 bases from the point of insert. In the case of the human genome this would generate 600 000 sequences that should be scattered approximately every 5 kb across the genome. These sequences can act as sequence-tagged connectors, or STCs, because they will allow any one BAC clone to be connected to about 30 others since a 150 kb insert 'divided ' by 5 kb will be represented in 30 BACs. In this way a physical map can be constructed. Each BAC clone can be fingerprinted using one restriction endonuclease to provide the insert size and detect artefactual clones by comparing the fingerprints with those of overlapping clones. It should be noted that this method has not yet been put into practice. Little (1996) expressed some doubts about its efficiency.

Assembling all the information

There are two kinds of maps: genetic maps and physical maps. The former kind is used to map genes of interest by analysis of the progeny derived from genetic crosses. RFLPs, CAPS, AFLPs and polymorphic STSs are examples of the molecular markers, which can be used in the generation of genetic maps. Physical maps are a prerequisite for genomic sequencing. Because only short lengths of DNA (typically 500 bases) can be sequenced in a single step, a physical map needs to have a much higher density of markers than a genetic map. Thus, for the purposes of sequencing the human genome, work is ongoing to generate a physical map in which 30 000 STS markers are placed on the genetic map of Dib *et al.* (1996) at an average density of 100 kb (Jordan & Collins 1996). Already preliminary maps were published (Hudson *et al.* 1995; Schuler *et al.* 1996). In the most recent version over 20 000 STSs had been mapped.

References

Adams, M.D, Dubnick, M., Kerlavage, A.R. *et al.* 1991. Complementary DNA sequencing: expressed sequence tags and human genome project [J]. Science, **252**: 1651-1656.

Archibald, A.L. 1994. Mapping of the pig genome [J]. Current opinion in Genetics and Development, 4: 359-400.

Archibald, A.L. 1995. The PiGMaP consortium linkage map of the pig (sus scrofa) [J]. Mammalian Genome, **6**: 157-175.

Barendse, W. 1994. A genetic linkage map of the bovine genome [J]. Nature Genetics, **6**: 227-235.

Bender, W., Spierer, P. and Hogness, D.S. 1983. Chromosome walking and jumping to isolate DNA from the Ace and rose loci and the bithorax complex in Drosophila melanogaster [J]. Journal of Molecular Biology, **168**:17-33.

Chumakov, I. 1992. Continuum of overlapping clones spanning the entire human chromosome 21q [J]. Nature, **359**: 380-387.

Chumakov, I.M. 1995. A YAC contig of the human genome [J]. Nature, **377**: S175-S298.

Dib, C. 1996. A comprehensive genetic map of the human genome based on 5264 microsatellites [J]. Nature, **380**: 152-154.

Dietrich, .W.F. 1994. A genetic map of the mouse with 4006 simple sequence length polymorphisms [J]. Nature Genetics, **7**: 220-225.

Dietrich, W.F. 1996. A comprehensive map of the mouse genome [J]. Nature, **380**:149-152.

Eggen, A. and Fries, R. 1995. An integrated cytogenetic and meiotic map of the bovine genome. Animal Genetics, **26**:215-236.

Foote, S., Vollrath, D., Hilton, A. *et al.* 1992. The human Y chromosome: overlapping DNA clones spanning the euchromatic region [J]. Science, **258**: 60-66.

Green, E.C. and Olson, M.V. 1990. Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping [J]. Science, **250**: 94-98.

Hudson, T.J. 1995. The STS-based map of the human genome [J]. Science, 270: 1945-1954.

Jacob, H.J. 1995. A genetic linkage map of the laboratory rat , Rattus norvegicus [J]. Nature Genetics, **9**: 63-69

Jordan, E. and Collins, F.S. 1996. A march of genetic maps [J]. Nature, **380**: 11-12.

Kurata, N. A. 1994. 300 kilobase interval genetic map of rice including 883 expressed sequences [J]. Nature Genetics, **8**: 365-372.

Little, P. 1996. Genome analysis [J]. Nature, **382**: 408

Makalowski W., Zhang .J. and Boguski M.S. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences [J]. Genome Research, **6**: 846-857.

Mazur, B.J. and Tingey, S.V. 1995. Genetic mapping and introgression of genes of agronomic importance [J]. Current Opinion in Biotechnology, **6**: 175-182.

Olson, M.V., Dutchik, J.E., Graham, M.Y., *et al.* 1989. A random-clone strategy for restriction mapping in yeast [J]. Proceedings of the National Academy of Science, USA, **83**: 7826-7830.

Postlethwait, J.H. 1994. A genetic linkage map for the zebrafish. Science, **264**: 699-704.

Reiter, R.S., Williams, J.G.K., Feldmann, K.A., *et al.* 1992. Global and local genome mapping in Arabidopsis thaliana by using recombinant inbread lines and random amplified polymorphic DNAs [J]. Proceedings of the National Academy of Sciences, USA, **89**: 1477-1481.

Schuler, G.D. 1996. A gene map of the human genome [J]. Science, **274**: 640-646.

Serikawa, T. 1992. Rat gene mapping using PCR-analyzed microsatellites [J]. Genetics, **131**: 701-721.

Smith, M.W., Holmsen, A.L., Wei, Y.H. *et al.* 1994. Genome sequence sampling: a strategy for high resolution sequence based physical mapping of complex genomes [J]. Nature Genetics, **7**: 40-47.

Tingey, S.V. and Del. Tufo, J.P. 1993. Genetic analysis with RAPD markers [J]. Plant Physiology, **101**: 349-352.

Venter, J.C., Smith, H.O. and Hood, L. 1996. A new strategy for genome sequencing [J]. Nature, **381**: 364-366.

Vos, P. 1995. AFLP: a new technique for DNA fingerprinting [J]. Nueleic Acids Research, **23**: 4407-4414.

Weissenbach, J., Gyapay, G., Dib, C. *et al.* 1992. A second generation linkage map of the human genome based on highly informative microsatellite loci [J]. Nature, **359**: 794-802.

Williams, J.G.K., Kubelik, A.R., Livak, K.J. *et al.* 1990. DNA polymorphisms amplified by arbitrary primers are usefule as genetic markers [J]. Nucleic Acids Research, **18**:6531-6535.